BioEdu
Berkala Ilmiah Pendidikan Biologi
Vol. 10 No. 1 Tahun 2021
Hal: 1-11
https://ejournal.unesa.ac.id/index.php/bioedu

DEVELOPMENT OF ASSESSMENT INSTRUMENT BASED ON HIGHER ORDER THINKING SKILLS OF RESPIRATORY SYSTEM OF GRADE XI OF SENIOR HIGH SCHOOL

Maria Ulfa
Biology Department, Faculty of Mathematics and Science, Universitas Negeri Surabaya
email: mariaulfa16030204097@mhs.unesa.ac.id

Nur Kuswanti
Biology Department, Faculty of Mathematics and Science, Universitas Negeri Surabaya
email: nurkuswanti@unesa.ac.id

Abstract

Higher order thinking skills (HOTS) is a thinking process at a higher level. It is not just memorizing facts, it involves mental activity in an effort to explore complex, reflective, and creative experiences. Indicators for measuring higher order thinking skills are analyzing, evaluating, and creating. These three indicators include critical thinking, creative thinking, problem solving, and decision-making skills. This research aimed to describe validity, reliability, and level of difficulty test items of higher order thinking skills of Respiratory System matter. This was development research referring to the ADDIE model (Analyze, Design, Development, Implementation, and Evaluation). The development of HOTS based assessment instrument was conducted at Biology Department – Universitas Negeri Surabaya and the trial activity was conducted at SMA Negeri 2 Sidoarjo. The theoretical validity was determined based on the results of the validation by matter expert and education expert using the validation sheet. The aspect of validity rated based on the aspects of concept, construction, language, and HOTS. The empirical validity was determined based on the result test items of HOTS on 40 students of class XI to identified empirical validity, reliability, and difficulty levels using the HOTS-based test items sheet. The results showed that the theoretical validity of the assessment instrument reached 98.19% and included in a very valid category. The empirical validity of 15 test items of HOTS were considered valid with the value of R (count) > 0.312. The reliability value of assessment instrument was 0.740 with a high level of consistency. The proportion of difficulty level was 60% with moderate category and 40% with difficult category. Based on the research results, it can be concluded that the HOTS-based assessment instrument of the Respiratory System was valid and reliable.

Keywords: higher order thinking skills, theoretical validity, empirical validity, reliability, level of difficulty, Respiratory System.

INTRODUCTION

The education of the 21st century with the implementation of the Curriculum 2013 requires students to have higher order thinking skills. It is not only remembering and understanding, restating, or referring without processing (Widana, 2017). The partnership of the 21st Century Skills identified that the achievement of the Curriculum 2013 in the 21st century emphasizes students to develop competitive skills that focus on the development of higher order thinking skills, such as critical thinking and problem solving, creativity and innovation, effective communication skills, media literacy, technology, and information (Basuki dan Hariyanto, 2016; Chalkiadaki, 2018).

Mapping basic competencies of biology subject matter of Curriculum 2013 identify that 23 of 35 competencies III target the achievement of high-level thinking (Kemendikbud, 2016). The analysis results of basic competency 3 of the Curriculum 2013 indicates that the cognitive dimensions spreaded are dominated by C-3 and C-4 levels, followed by 15% achievement of C-2 level (Herlanti, 2015). One of the biology matter that targets the achievement of high-level thinking is the Respiratory System. The basic competence III of class XI of senior high school as describes in Basic Competence 3.8 (KD 3.8.) requires students to be able to think at a high level because the competence targeted is at the cognitive dimension of C-4 (Analyze). Lissa et al. (2012) explained that the Respiratory System is a subject matter of biology relating to the physiology of the body and respiratory organs in living things. Besides, this also relates to daily life, such as events in the community about the effects of air pollution and the influence of smoking habits on respiratory health or disorders of the respiratory tracts. So that, Setiawati et al. (2019) explained that the Ministry of Education and Culture through the Directorate General of Teachers and Educational Staff conducts a Learning Competency Improvement program as an effort to improve the quality of sustainable learning in improving the quality of students. The program is about the HOTS-Based Assessment.

Assessment is a process of collecting data or information to measure the results of student learning, by educators, education units, or governments (Kemendikbud, 2016). Assessment can also be used to evaluate the learning process, diagnose students' weaknesses in certain subjects, and develop a progress report on the results of learning (Widana, 2017). HOTS-based assessment functions to stimulate learning access in a complex, reflective, and creative experience to get knowledge in analysis, synthesis, and evaluation thinking levels (Abosalem, 2016). A HOTS-based assessment instrument needs to be developed as a reference to face the advances of science and technology. According to Richmond (2017), good thinking skills can be an asset for Asian students to handle complex problems. Habituating to use higher order thinking, Indonesian students should be trained to solve problems based on real situations in their daily life.

The dimensions of cognitive processes of Bloom's Taxonomy refined by Anderson & Krathwohl are classified into three levels, that are Level-1 (C-1 and C-2), Level-2 (C-3), and Level-3 (C-4, C-5, and C-6) (Widana, 2017). Indicators for measuring higher order thinking skills are analyzing, evaluating, and creating. These three indicators include critical thinking, creative thinking, problem solving, and decision-making skills. Critical thinking is the ability to identify, connect, analyze, solve problems, and draw conclusions based on data, evidence, rational and logical reasons (Trilling and Fadel, 2009; Palinussa, 2015; Yuliani and Saragih, 2015). Individuals who think critically always use their knowledge and experience to identify and analyze new things, so that they can justify or make decisions (Susilo, et al., 2019). Creative thinking is the ability to find new or different ideas. By this thinking, individuals can create various innovations to solve various problems related to their daily life (Setiawati et al., 2019).

The quality of test items made by teachers in Indonesia are still low (Rofiah et al., 2013). Based on the preliminary study, the results of interviews and analysis of documents of test items in a state high school of Sidoarjo region, it showed that the tests made by the teacher were in range of level-1 and level-2. These were supported by the teacher's test document for daily, midterm, and final exam assessments that mostly were still in the dimensions of cognitive processes of C-1, C-2, and C-3. Accordingly, the test items made by the biology teacher of the high school cannot be used to measure students' higher order thinking skills, so that the ability of students is also not yet measured. In addition, the results of cognitive tests of biology subjects were mostly at LOTS levels, including the levels of C-1 (92%) and C-2 (83%). However, the results of the dimensions of cognitive processes of C-3, C-4, C-5, and C-6, got values of 22%, 22%, 17%, and 8%.

A research conducted by Masruroh et al. (2012) identified similar results that the odd semester test items of biology class X in Kebumen region measured cognitive levels of C-1 and C-2. The results of other study conducted by Putri et al. (2018) also showed that the test items on the aspect of higher order thinking for grade X of senior high school were still in the dimensions of cognitive processes of C-1 (47%), C-2 (47%), and C-3 (6 %). Binethara et al. (2017) also found that the test items of midterm and final exam assessments of state high schools in Gadingrejo region only measured cognitive levels of C-1, C-2, and C-3 with the knowledge dimensions of factual, conceptual, and procedural.

Higher order thinking skill based assessments can be in the form of subjective and objective tests. The subjective test is a form of essay test, which students can describe answers using their own sentences. The objective tests are tests having true-false, multiple choices, completion, and matching answers (Asrul et al., 2015). In identifying students' competence of the higher order thinking skills, it should accommodate both such forms, because the dimension of the C-6 (Creating) cognitive process tests is more suitable to use the essay test, but the dimensions of C-4 and C-5 cognitive processes can apply the essay or multiple choice forms (Asrul et al., 2015).

Before it is given to students, assessment instrument that have been created should be reviewed or analyzed, so their quality can be identified (Hasanah et al., 2016). Using instrument analysis aims to study and evaluate test items in measuring or identifying students' achievement of competencies (Masruroh et al., 2012). It can be done qualitatively and quantitatively. Using qualitative or theoretical analysis', they are reviewed based on the aspects of the matter, language, and construction. Using quantitative or empirical analysis' they can be reviewed based on their item's validity, reliability, and level of difficulty (Arikunto, 2015).

This research aimed to describe validity, reliability, and levels of difficulty of test items of higher order thinking skills of the Respiratory System matter.

METHODS

This study was a development research being to develop test items of Higher Order Thinking Skills (HOTS) of Respiratory System matter. This development research referred to the ADDIE model consisting of five stages, those are Analyze, Design, Development, Implementation, and Evaluation (Pribadi, 2016).

The analysis stage was carried out based on three aspects, i.e. curriculum analysis, student analysis, and concept analysis. The second was the stage of the creating test items of HOTS. The test items of Higher Order Thinking Skills (HOTS) were presented in the form of multiple choice and essay. They were developed in the cognitive dimensions of C-4, C-5, and C-6. The aspects of HOTS used were critical thinking, creative thinking, problem solving, and decision-making. The preparation of test items were done by detailing the basic competence into several indicators. The test items specification of HOTS of Respiratory System matter are presented in Table 1.

Table 1. The Test Item Specification of HOTS of Respiratory System Matter

Basic Competencies	Indicator of Achieving Competence	Indicators of HOTS	Aspect of HOTS	Cognitive Level	Number	Instrument Form
3.8 Analyzing the relationship between the structure of the tissue in organs of the respiratory system in relation to bioprocess and functional disorders that can occur in human respiratory system.	Concluding factors affecting the respiratory frequency.	Determining decision based on information or problem given.	Decision-making	C-4	1	Multiple Choice
	Analyzing the relationship between the respiratory frequency and the pulse.	Identifying, processing, analyzing, and connecting information to concepts, theories, and opinions.	Critical Thinking	C-4	2	Multiple Choice
	Predicting the results of an experiment of breathing air composition.	Solving problems based on data or information given.	Problem Solving	C-5	3	Multiple Choice
	Analyzing the volume and capacity of human lungs.	Identifying, processing, analyzing, and connecting information to concepts, theories, and opinions.	Critical Thinking	C-4	4	Multiple Choice
		Determining decision based on information or problem given.	Decision-making	C-4	5	Multiple Choice
	Analyzing the relationship between factors and the volume and capacity of human lungs.	Identifying, processing, analyzing, and connecting information to concepts, theories, and opinions.	Critical Thinking	C-4	6, 9	Multiple Choice
	Analyzing the influence of factors causing abnormalities or functional disorders that can occur in the human respiratory system.	Identifying, processing, analyzing, and connecting information to concepts, theories, and opinions.	Critical Thinking	C-4	7, 10, 11, 12	Multiple Choice
	Identifying disorders of the respiratory system based on lungs' volume and capacity graphs.			C-4	8	Multiple Choice
	Formulating solution to prevent disorders of human respiratory system.	Solving problems based on data or information given.	Problem Solving	C-5	13,14	Essay
	Designing experimental devise to prove the effect of cigarette smoke's composition on body health.	Solving problems with ideas.	Creative Thinking	C-6	15	Essay

Then, the development stage was conducted through some substages. In this stage, the instrument as draft 1 were reviewed then revised. The revision produced draft 2. The last was validated by experts using a validation sheet. After that, draft 3 was tested to students. The development of HOTS based assessment instrument was conducted at Biology Department – Universitas Negeri Surabaya and the trial activity was done at SMA Negeri 2 Sidoarjo with 40 students of class XI. The implementation stage obtained values of empirical validity, reliability, and difficulty level.

Data gained were analysed qualitatively and quantitatively. The qualitative analysis was conducted through reviewing tests items to experts to determine the theoretical validity of assessment instrument. The theoretical validity of HOTS-based assessment instrument was calculated using the following formula.

$$P = \frac{f}{N} x\ 100\%$$

Notes:

P = percentage of validity (%)
f = total of test items marked on each aspect (√)
N = total of all test items

The percentage of validity were interpreted based on the criteria as shown in Table 2.

Table 2. Interpretation of Validity

Validity (%)	**Interpretation of Validity**
$81.50 \leq P \leq 100.00$	Very Valid
$62.75 \leq P \leq 81.49$	Valid
$44.00 \leq P \leq 62.74$	Valid Enough
$25.00 \leq P \leq 43.99$	Less Valid
$00.00 \leq P \leq 24.99$	Not Valid

Source: Riduwan (2012)

The assessment instrument being theoretically valid, then tested on 40 students of class XI to identify its empirical validity, reliability, and difficulty level. The empirical validity of test items was analyzed using SPSS 23 software by Product Moment correlation formula as follow.

$$g_{pbi} = \frac{Mp - Mt}{St}\sqrt{\frac{p}{q}}$$

Notes:

g_{pbi} = biserial correlation coefficient
M_p = an average score of subject that responds correctly to each test item
M_t = average total score
S_t = standard deviation of the total score
p = proportion of students answering correctly $(p = \frac{\text{total students who answered correctly}}{\text{total number of students}})$
q = proportion of students answering incorrectly $(q = 1 - p)$

Source: (Arikunto, 2015)

Analyzing the reliability of test items used the Cronbach Alpha formula being applied in the Alpha model of SPSS 23 software as below.

$$r_{11} = \left(\frac{n}{(n-1)}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right)$$

Notes:

r_{11} = reliability of instrument
$\sum \sigma_i^2$ = total of score variances of each item
n = total of items
σ_t^2 = total variance

The calculation results were interpreted based on the criteria as shown in Table 3.

Table 3. Interpretation of Reliability

Interval of Reliability	**Interpretation of Reliability**
$0.80 < r_{11} \leq 1.00$	Very reliable
$0.60 < r_{11} \leq 0.80$	Reliable
$0.40 < r_{11} \leq 0.60$	Reliable Enough
$0.20 < r_{11} \leq 0.40$	Less Reliable
$0.00 \leq r_{11} \leq 0.20$	Not Reliable

Source: Sugiyono (2015)

The difficulty level of each test item was determined using the following formula.

$$P = \left(\frac{B}{Js}\right) x\ 100\%$$

Notes:

P = level of difficulty
B = number of students who answered correctly
Js = total number of the students

Source: (Arikunto, 2015)

The test result of each item was calculated using the following formula.

$$P = \frac{\sum x}{Sm\ x\ N}$$

Notes:

P = level of difficulty
$\sum x$ = total score obtained
Sm = maximum score
N = total number of students

Source: (Arikunto, 2015)

The calculation result was interpreted based on the criteria shown in Table 4.

Table 4. Interpretation of Difficulty Level

Interval of Difficulty Level	**Interpretation of Difficulty Level**
$0.00 \leq P \leq 0.30$	Difficult
$0.30 < P \leq 0.70$	Moderate
$0.70 < P \leq 1.00$	Easy

Source: Arikunto (2015)

RESULTS AND DISCUSSION

This research aimed to describe the validity, reliability, and level difficulty of test items of higher order thinking skills (HOTS) instrument of Respiratory System matter. This development research referred to the ADDIE model (Analyze, Design, Development, Implementation, and Evaluation).

Analysis was the first stage of the development of test items of HOTS. This stage was carried out by identification of basic competencies to develop indicators. Formulating indicators were done according to the basic competence relating to Respiratory System matter that should be achieved through higher order thinking skills. The indicators were developed according to indicators of HOTS, aspects of HOTS, and dimensions of cognitive processes. The results of this stage are organized in Table 1.

The second stage was creating test items of HOTS that reffered to indicators formulated (Table 1). There were 15 test items of HOTS developed, i.e. 12 multiple choice and 3 essay type test items. Two example of test items of HOTS are presented in **Figure 1.**

Multiple Choice Type

Para nelayan tradisional merupakan penyelam bebas yang bisa menyelam hingga di kedalaman laut tertentu tanpa menggunakan bantuan tabung oksigen. Mereka bisa menahan napas dan bertahan dalam waktu yang lama ketika berada di dalam air. Aktivitas tersebut berkaitan dengan volume udara pernapasan. Hal apa yang menyebabkan nelayan dapat bertahan pada kondisi tersebut?

A. Volume tidal yang sangat besar.
B. Volume residu yang sangat besar.
C. Kapasitas vital paru-paru yang besar.
D. Volume cadangan inspirasi yang besar.
E. Volume cadangan ekspirasi yang besar.

Essay Type

Pada tahun 2019, beberapa daerah di Indonesia terjadi peristiwa kabut asap akibat kebakaran hutan dan lahan. Paparan kabut asap tidak hanya mengganggu aktivitas masyarakat di sekitarnya, tetapi juga berdampak bagi kesehatan, terutama pada saluran pernapasan. Berbagai macam penyakit saluran pernapasan yang muncul akibat kebakaran hutan, antara lain asma, bronkitis, pneumonia, hingga penyakit paru obstruktif kronis. Kebakaran hutan dan lahan menimbulkan masalah kesehatan yang dapat memiliki efek jangka pendek atau jangka panjang bagi manusia, mulai dari bayi, anak-anak, remaja, dewasa, hingga para lansia.

Berdasarkan ulasan kasus di atas, siapa yang paling berisiko mengalami gangguan kesehatan akibat kabut asap? Sertakan dengan alasan! Jika tempat tinggal Anda berada dalam kawasan lingkungan yang rawan mengalami kabut asap, maka bagaimana cara Anda untuk menjaga kesehatan sistem pernapasan?

Jawab : ______________________________
____
____

Figure 1. Example of Test Items of HOTS

Creating test items used stimulus that encouraged students to think. Each test item of HOTS developed referred to the dimension of cognitive processes of C-4, C-5, or C-6 and one of HOTS aspects, including critical thinking, creative thinking, problem solving, or decision-making.

The development stage was conducted through several substages. Firstly, an initial instrument was created and named as draft 1. The draft 1 were reviewed then revised to produce draft 2. The last was validated by experts using a validation sheet. In this stage, the items of HOTS were reviewed and validated based on the aspects of matter/concept, construction, language, and HOTS. It was done to determine whether the items were good or not (Rahmani, 2015). The theoretical validity values of the HOTS-based assessment instrument of multiple choice and essay types are presented in Table 5. and Table 6.

Table 5. The Theoretical Validity of Multiple Choice Test

No	Assessment Aspects	Average (%)		Validity (%)	Interpretation
		V1	V2		
A.	**Matter/Concept**				
1	Each test item refers to an indicator.	100	100	98.96	Very Valid
2	Each test item refers to the correct concepts.	91.7	100		
3	The scope of each test item is clear and there is only one correct answer.	100	100		
4	The content of matter refers to levels of school and class.	100	100		
B.	**Construction**				
1	The instructions of each test item are easy to be understood.	83.3	83.3	96.29	Very Valid
2	There is scoring guideline.	100	100		
3	Each test item does not contain double answers.	100	91.7		
4	Each test item does not depend on other answers of the other test items.	100	83.3		
5	Each test item does not give any clue of the correct answer option.	100	100		
6	The options of each test item are relatively equal.	100	100		
7	The options of each test item are homogeneous and logical.	100	100		
8	The option of each test item does not contain statement that all answers are	100	100		

No	Assessment Aspects	Average (%)		Validity (%)	Interpretation
		V1	V2		
	correct or incorrect.				
9	Figures, graphs, tables, or diagrams are clear and functional.	100	91.7		
C.	**Language**				
1	Grammar and spelling refer to *Ejaan Bahasa Indonesia* (EBI).	100	100	90.27	Very Valid
2	Using communicative language and being easy to be understood.	83.3	83.3		
3	Each sentence does not cause double interpretations or misunderstanding.	91.7	83.3		
D.	**Higher Order Thinking Skills**				
1	Each test item refers to the dimension of cognitive processes of C-4, C-5, or C-6.	100	100	100	Very Valid
2	Each test item refers to one of HOTS aspects, including critical thinking, creative thinking, problem solving, or decision-making.	100	100		
3	Each test item uses stimulus that encourages students to think.	100	100		
Average of all aspects (%)				**96.38**	Very Valid

Based on Table 5., it can be seen that the overall theoretical validity of the multiple choice test reached 96.38% with a very valid category.

Table 6. The Theoretical Validity of Essay Test

No	Assessment Aspects	Average (%)		Validity (%)	Interpretation
		V1	V2		
A.	**Matter/Concept**				
1	Each test item refers to an indicator.	100	100	100	Very Valid
2	Each test item refers to the correct concepts.	100	100		
3	The scope of each test item is clear.	100	100		
4	The content of matter refers to levels of school and class.	100	100		
B.	**Construction**				
1	Each test item uses question or instruction that require student reasoning.	100	100	100	Very Valid
2	Each test item uses a clear guidance to answer questions or do any instruction.	100	100		
3	There is scoring guideline.	100	100		
4	Each test item does not depend on other answer of the other test items.	100	100		
C.	**Language**				
1	Grammar and spelling refer to EBI.	100	100	100	Very Valid
2	Using communicative language and being Easy to be understood.	100	100		
3	Each sentence does not cause double interpretations or misunderstanding.	100	100		
D.	**Higher Order Thinking Skills**				
1	Each test item refers to the dimension of cognitive processes of C-4, C-5, or C-6.	100	100	100	Very Valid
2	Each test item refers to one of HOTS aspects, including critical thinking, creative thinking, problem solving, or decision-making.	100	100		
3	Each test item uses stimulus that encourages students to think.	100	100		
Average of all aspects (%)				**100**	Very Valid

Based on Table 6., it can be seen that the overall theoretical validity of the essay test reached 100% with a very valid category. The results of the overall theoretical validity of the multiple choice and essay types are presented in Table 7.

Table 7. The Theoretical Validity of Multiple Choice and Essay Test Item

Type of Tests	Assessment Aspects (%)				$\bar{x}$ (%)
	Matter	Construction	Language	HOTS	
Multiple Choice	98.96	96.29	90.27	100	96.38
Essay	100	100	100	100	100
Average of all test items (%)					**98.19**
Interpretation					**Very Valid**

Based on Table 7., the theoretical validity of test items, multiple choice and essay types, were determined based on the aspects of matter/concept, construction, language, and HOTS. Accordingly, the validity reached 98.19% with a very valid category. However, the test items of HOTS still need improvement referred to suggestions from experts, such as shown in Table 8.

Table 8. Suggestions for test items of HOTS

No	Suggestions	
	Before	After
1	Diketahui ada dua sampel udara dari sumber yang berbeda yang masing-masing ditampung di dalam **tabung tersendiri.**	Perhatikan gambar eksperimen berikut. Dua sampel udara dari sumber yang berbeda ditampung di dalam **tabung berbeda.**
2	Suatu hari, **Pak Ali** merasakan batuk, nyeri pada bagian dada, dan sesak napas.	Suatu hari, ada **seorang pasien** yang merasakan batuk, nyeri pada bagian dada, dan sesak napas.
3	Suatu hari, Polisi menemukan 6 dari 11 jiwa meninggal dunia dalam kurun waktu ± 12 jam. Mereka ditemukan di kamar mandi berukuran 1,5 m x 1,5 m. Berdasarkan hasil identifikasi, **korban yang meninggal dunia terlihat pucat ungu kebiruan pada tubuhnya.**	Suatu hari, ada konser musik dan launching grup band di gedung AACC Kota Bandung. Saat konser berlangsung, ditemukan banyak korban jiwa yang **jatuh pingsan**. Jumlah penonton yang hadir pun lebih banyak daripada kapasitas gedung yang hanya dapat menampung 400 jiwa.

Based on Table 8., it can be concluded that the suggestions of experts are about 1) using communicative language referred to the level of cognitive development of students, 2) avoiding to use the name of any person, and 3) the stimulus of questions must be as short as possible and avoid horrible stories in creating item test.

Theoretical validation is an instrument feasibility assessment procedure for generating agreements among experts through qualitative assessments (Azwar, 2016). A theoretical validity can be determined based on the accuracy and sensibility of the measurement results by experts using a validation sheet (Arikunto, 2015).

The validity of assessment instrument developed is categorized based on three aspects, i.e. matter/concept, construction, and language. The aspect of matter relates to scientific substance manifested in the test items and cognitive levels (Asrul et al., 2015). The aspects of matter got validation results with very valid interpretations, i.e. 98.96% for multiple choice type and 100% for essay type. These because the tests' matter aspect already referred to the matter/concept contained in the basic competency referred and indicators developed. Moreover, the content of matter referred to the level of school and class. Thus a test has good matter validity if it is comprehensive, representative, and relevant. A comprehensive test refers to the content or subject matter being based on basic competencies. In addition, a representative tests is a test items that can be used as measuring assessment instrument. A relevant tests is a test item referring to the correct concept (Arikunto, 2015).

The aspects of construction relates to the rules of writing tests. In this study, the aspects of construction got very valid categories, i.e. 96.29% for multiple choice tests and 100% for essay tests. Some rules have to be considered in writing both test types. They should have clear instructions, be easy to be understood, be clear, and have scoring guidelines (Arikunto, 2015). A test has good construction validity if it refers to the rules of writing test, so that using the test can evaluate thinking skill of each student (Asrul et al., 2015).

The aspects of language referred to general Guidelines for Indonesian Spelling (*Pedoman Umum Ejaan Bahasa Indonesia*/PUEBI). The aspects of language got very valid interpretations. Multiple choice tests reached 90.27% and essay tests reached 100%. These results indicate that the HOTS-based assessment instrument referred the PUEBI rules. It was shown by its sentences that had no double interpretations, and were communicative referring to the level of cognitive development of students. Creating assessment instrument needs to pay attention on aspects of language, because using good and correct language can facilitate students to understand the meaning of test sentences. So that the students have no difficulty to understand the tests (Arikunto, 2015).

The test items of Higher Order Thinking Skills (HOTS) were very valid. It indicates that the test items referred to the dimension of cognitive processes of C-4, C-5, and C-6. These three indicators include critical thinking, creative thinking, problem solving, and decision-making skills. Moreover, the test items exert stimulus that encourages students to think. HOTS's test items in the assessment would train the students to hone their abilities and skills in accordance with the demands of the 21st century competencies (Sulaiman et al., 2017). Through activities test items HOTS, students will solve various problems in daily life, so that they can improve their skill to think critically, creatively, and innovatively (Richmond, 2017).

The assessment instrument being theoretically valid was tested on 40 students of class XI to identify its empirical validity, reliability, and difficulty levels. Identification of the empirical validity aimed to determine wether each test item developed was either valid or invalid. The empirical validity was analyzed by correlating the scores of items with their total score using the correlation of Product Moment of SPSS 23 software. The empirical validity values of test items HOTS are shown in Table 9.

Table 9. The Empirical Validity of Test Item HOTS of Respiratory System

No of Test	R_{table}	R_{count}	Category
1	0.312	0.420	Valid
2	0.312	0.435	Valid
3	0.312	0.343	Valid
4	0.312	0.313	Valid
5	0.312	0.429	Valid

6	0.312	0.322	Valid
7	0.312	0.466	Valid
8	0.312	0.327	Valid
9	0.312	0.463	Valid
10	0.312	0.651	Valid
11	0.312	0.638	Valid
12	0.312	0.349	Valid
13	0.312	0.626	Valid
14	0.312	0.715	Valid
15	0.312	0.707	Valid

Based on Table 9. it can be seen that the empirical validity of 15 HOTS test items were considered valid empirically, because of the values of R (count) being higher than R (table) (R_{count}> 0,312). Each value of R (table) was obtained from the Pearson R Correlation table at a significant rate of 5% (N = 40. A research conducted by Ikhsan (2017) showed 10 test items developed were considered valid (R_{count} > 0.58). It is similar to the result of recent study. An assessment instrument that is considered valid by experts does not ensure that it will be valid after being tested testing on students. Arikunto (2015) explained that if the test items are empirically valid, it means that the instrument can measure and exhibit its next results precisely and accurately. Widoyoko (2016) also explained that if each test item created is valid, the assessment instrument is also valid. Thus, a high validity assessment instrument can measure the achievement of objectives. In this case, the instrument provides precise and accurate measuring results (Sugiyono, 2015).

Identifying reliability aimed to determine the consistency or persistence of an assessment instrument. In this study, the calculation of reliability values used the Alpha model of SPSS 23 software by applying the Cronbach Alpha formula. The result is shown in Table 10.

Table 10. Reliability of HOTS-based Assessment Instrument of Respiratory System

Reliability Value	Category
0.740	High/Reliable

Based on Table 10., the reliability value of assessment instrument was 0.740 with a high level of consistency shown by $r_{11} \geq 0.61$. A conducted by Putri (2017) showed that the developed HOT assessment instrument got reliability value of 0.58 with a moderate level of consistency. Ghozali (2011) explained that if the value is > 0.60, the instrument is reliable. However if the Cronbach Alpha value is < 0.60, the instrument is not reliable. A reliable test can result a consistent score. The result is relatively fixed and unchanged even if it is tested at different times (Arikunto, 2015). Some factors influence the value of reliability, either directly or indirectly. The direct factors include the time of implementation, the length of test item, the test difficulty index, the distribution of values, and the objectivity of scoring. On the other hand, direct factors are the clarity of implementation instructions, supervision, and environmental conditions (Retnawati et al., 2018).

Identifying the difficulty level of test items aimed to determine the distribution of HOTS test items. The calculation results of each item developed are presented in Table 11.

Table 11. The Difficulty Level of Test Items HOTS of Respiratory System

No of Test	Level of Difficulty	Category
Test Items of Multiple Choice		
1	0.60	Moderate
2	0.70	Moderate
3	0.33	Moderate
4	0.30	Difficult
5	0.28	Difficult
6	0.23	Difficult
7	0.58	Moderate
8	0.28	Difficult
9	0.48	Moderate
10	0.28	Difficult
No of Test	**Level of Difficulty**	**Category**
11	0.70	Moderate
12	0.58	Moderate
Test Items of Essay		
13	0.49	Moderate
14	0.41	Moderate
15	0.19	Difficult

Based on Table 11. it can be seen that the 15 items have various difficulty levels, consisting of 9 items of moderate category and 6 items of difficult categories. The proportion of the levels is presented in as diagram as shown in Figure 2.

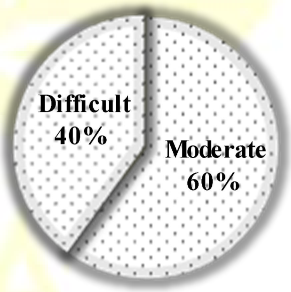

Figure 2. Proportion of Difficulty Levels

Based on Figure 2. the difficulty in HOTS-based assessment instrument reach 60% for moderate category and 40% for difficult category. There is no definite proportion of difficulty levels, because the basis of its determination is the purpose of using the assessment instrument (Arikunto, 2015).

The level of difficulty is one of the parameters to analyze test items in determining students' skills (Retnawati et al., 2018). A good test is being not too easy or not too difficult. A too easy test can not encourage students to improve their efforts in solving problems. Conversely, a too difficult test can cause students become discouraged and be unable to answer question, because it is beyond their thinking (Arikunto, 2015).

There was no easy level of HOTS-based test item, because each test can be answered correctly by few students. A test item belongs to an easy category if it can be answered correctly by most students involved, of both the upper and lower groups (Arikunto, 2015). Thus, the easy test items can not encourage students to think. Contrary, the difficult instrument developed with the characteristic of HOTS provides items contextual-based that encourage students to think in solving problems relating to daily life.

The instrument developed contain nine items included in moderate category of difficulty levels, i.e. tests number 1, 2, 3, 7, 9, 11, 12, 13, and 14. According to the provision of test item creation, a good test is not too easy or too difficult. In addition, the instrument also contain six difficult items, i.e. test number 4, 5, 6, 8, 10, and 15. Rokhyati (2011) explained that there were some problems of test items including difficult category, e.g. 1) the question has two or more correct answers, 2) the subject matter targeted has not been learnt or not been completed to be learnt by students, so that the minimum competence has not been mastered, 3) the test form used is not suitable to measure the achievement of material mastery, and 4) questions or sentences used are too long.

In general, the difficulty level only indicates that the question items include in difficult, moderate, or easy categories for a specific group of students. In assessing activities, students were divided into two groups, upper and lower groups. If the item can be done both groups, it includes in the easy category. If the test item can only be done by the upper group, it includes in the difficult category. It does not mean that the test items that are too easy or too difficult can not be used. However, it depends on the purpose of using the assessment instrument. Although they do not provide much information about the problem or the ability of students (Arikunto, 2015).

The HOTS-based assessment instrument of the Respiratory System matter developed contain three higher order thinking indicators, consisting of 11 items of the cognitive dimension of C-4 (Analyze), 3 items of the cognitive dimension of C-5 (Evaluate), and 1 item of the cognitive dimension of C-6 (Create). These three indicators include critical thinking, creative thinking, problem solving, and decision-making skills. The HOTS-based assessment instrument of the Respiratory System was considered valid and reliable. Based on the results of the theoretical validity, empirical validity, and reliability. Azwar (2016) explained that a good instrument has valid and reliable characteristics. A valid test will serve as a measuring instrument in learning. In addition, the use of test depends on measuring purpose.

CONCLUSION

Based on the research results and discussion, it can be concluded that the HOTS-based assessment instrument of the Respiratory System was valid and reliable. The theoretical validity of assessment instrument developed in the aspects of matter/concept, construction, language, and Higher Order Thinking Skills (HOTS) reached 98.19% with a very valid category. The 15 test items of HOTS were considered empirically valid shown by R (count) > 0.312 of each item. The reliability value of the assessment instrument developed was 0.740 with a high level of consistency (reliable). And, the test items developed have two levels of difficulty, 60% of them are categorized as moderate level and 40% of them are difficult.

SUGGESTION

Creating assessment instrument have to be based on the achievement targets of basic competencies. Before being given to students, a teacher should make an assessment instrument referring to the pointing to the competency and dimensions of cognitive processes. This research is recommended to be followed up with topics below.

1. Research on profiles of higher order thinking skills of students.
2. Research on the development of high-level thinking handouts
3. Research on the development of HOTS-based assessment instrument on other subject matters.

REFERENCES

Abosalem, Y. 2016. Assessment Techniques and Students Higher-Order Thinking Skills. *International Journal of Secondary Education.* Vol. 4 (1), hal. 1-11.

Arikunto, S. 2015. *Dasar-Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara.

Asrul, Ananda, R., dan Rosnita. 2015. *Evaluasi Pembelajaran*. Bandung: Citapustaka Media.

Azwar, S. 2016. *Konstruksi Tes Kemampuan Kognitif.* Yogyakarta: Pustaka Belajar.

Basuki, I., dan Hariyanto. 2016. *Asesmen Pembelajaran.* Bandung: PT Remaja Rosdakarya.

Binethara, P., Achmad, A., dan Yolida, B. 2017. Identifikasi UTS dan UAS Mata Pelajaran Biologi Berdasarkan Taksonomi Bloom Revisi

Anderson. *Jurnal Pendidikan Biologi FKIP Universitas Lampung*. Vol 4 (1), hal.1-12.

Chalkiadaki, A. 2018. A Systematic Literature Review of 21st Century Skills and Competencies in Primary Education. *International Journal of Instruction*. Vol. 11 (3), hal. 1-16.

Ghozali, I. 2011. *Aplikasi Analisis Multivariate dengan Program IBM SPSS 20*. Semarang: Universitas Diponegoro.

Hasanah, Annisah L., Subali, B., dan Mariyan, S. 2016. Analisis Item Ujian Akhir Semester (UAS) Genap Mata Pelajaran Biologi Kelas X Tahun Ajaran 2014/2015 di Sekolah Menengah Atas yang Mengimplementasikan Kurikulum 2013 di Kabupaten Sleman. *Jurnal Pendidikan Biologi*. Vol. 5 (4), hal. 7-26.

Herlanti, Y. 2015. Kesadaran Metakognitif dan Pengetahuan Metakognitif Peserta Didik SMA dalam Mempersiapkan Ketercapaian Standar Kelulusan pada Kurikulum 2013. *Jurnal Cakrawala Pendidikan*. Th.XXXIV, No.3, hal. 357-367.

Ikhsan, M. 2017. Uji Validitas dan Reliabilitas Instrumen Tes Uraian Berbasis *Higher Order Thinking Skills* (HOTS) Pada Materi Hidrolisis Garam Untuk Siswa SMA/MA. *Jurnal Pendidikan Kimia*. Vol.6 (3), hal. 1-5.

Kemendikbud. 2016. *Permendikbud No.23 tentang Standar Penilaian Pendidikan.* Jakarta: Kementerian Pendidikan dan Kebudayaan.

Lissa, Prasetyo, A. P., dan Indriyanti, D. R. 2012. Pengembangan Instrumen Penilaian Keterampilan Berpikir Tingkat Tinggi Materi Sistem Respirasi dan Ekskresi. *Lembaran Ilmu Kependidikan*. Vol. 41 (1), hal. 27-32.

Masruroh, Rudyatmi, E., dan Ridlo, S. 2012. Analisis Soal Ulangan Semester Gasal Biologi Kelas X Di Kecamatan Petanahan Kebumen. *Journal of Biology Education Unnes*. Vol 1 (2), hal. 116-121.

Palinussa, A. L. 2015. Students' Critical Mathematical Thinking Skills and Character: Experiments for Junior High School Students through Realistic Mathematics Education Culture-Based. *IndoMS. J.M.E*. Vol. 4 (1), hal. 75-94.

Pribadi, B. A. 2016. *Desain dan Pengembangan Program Pelatihan Berbasis Kompetensi Implementasi Model ADDIE*. Jakarta: Prenada Media Group.

Putri, B. A. Y. 2017. Validitas Empiris Butir Soal *High Order Thinking* (HOT) Berbasis *Computer Based Test* (CBT) pada Sub Materi Sistem Indera Siswa Kelas XI SMA. *Jurnal Bioedu*. Vol. 6 (3), hal 353-359.

Putri, et al. 2018. Analisis Aspek Kemampuan Berpikir Tingkat Tinggi pada Instrumen Penilaian Materi Protista Untuk Peserta Didik SMA/MA Kelas X. *Jurnal Biodik*. Vol. 4 (1), hal 2460-2612.

Rahmani, M. 2015. Analisis Kualitas Butir Soal Buatan Guru Biologi Kelas X SMA Negeri 1 Tanah Pinoh. *Artikel Penelitian Pendidikan*. Pontianak: Universitas Tanjungpura.

Retnawati, H., et al. 2018. Teachers; Knowledge About Higher-Order Thinking Skills and Its Learning Strategy. *Problesms of Education The 21st Century*. Vol 76 (2), hal. 1-7.

Richmond, J. 2017. Bringing Critical Thinking to The Education Of Developing Country Professionals. *International Education Journal*. Vol. 8 (1), hal. 1-29.

Riduwan. 2012. *Skala Pengukuran Variabel-Variabel Penelitian*. Bandung: Alfabeta.

Rofiah, E., Aminah, N. S., dan Ekawati, E. Y. 2013. Penyusunan Instrumen Tes Kemampuan Berpikir Tingkat Tinggi Fisika pada Siswa SMP. *Jurnal Pendidikan Fisika*. Vol. 1 (2), hal. 17-22.

Rokhyati. 2011. Karakteristik Secara Kualitatif dan Kuantitatif Soal Ulangan Akhir Semester Genap Bahasa Indonesia Kelas XII SMA Negeri di Kabupaten Purbalingga Tahun Pelajaran 2010/2011. Skripsi. Universitas Negeri Yogyakarta: diterbitkan di https://eprints.uny.ac.id (diakses pada tanggal 13 Juli 2019).

Setiawati, W., Asmira, O., Ariyana, Y., Bestary, R., dan Pudjiastuti, A. 2019. *Buku Penilaian Berorientasi Higher Order Thinking Skills.* Jakarta: Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan.

Sugiyono. 2015. *Metode Penelitian Pendidikan (Pendekatan Kuantitatif, Kualitatif, dan R&D).* Bandung: CV Alfabeta.

Sulaiman, T., et al. 2017. Implementation of Higher Order Thinking Skills in Teaching Science: A Case Study in Malaysia. *International Research Journal of Education and Science (IRJES)*. Vol. 1 (1), hal. 1-3.

Susilo, E. F., Abdurrahman A., and Malan L. 2019. The Ability to Understand Questions of Writing Scientific Works based on Higher Order

Thinking Skills (HOTS). *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal*. Vol. 2 (2), hal. 360-371.

Trilling, B., and Fadel, C. 2009. *21st Century Skills: Learning for Life in Our Times.* San Fransisco: Jossey-Bass.

Widana, I. W. 2017. *Modul Penyusunan Soal Higher Order Thinking Skills* (*HOTS*). Jakarta: Direktorat Jenderal Pendidikan Dasar dan Menengah Departemen Pendidikan dan Kebudayaan.

Widoyoko, E. P. 2016. *Evaluasi Program Pembelajaran*. Yogyakarta: Pustaka Belajar.

Yuliani, K., and Saragih, S. 2015. The Development of Learning Devices Based Guided Discovery Model to Improve Understanding Concept and Critical Thinking Mathematically Ability of Students at Islamic Junior High School of Medan. *Journal of Education and Practice*. Vol. 6 (24), hal. 116-129.